휴머노이드 로봇을 위한 원거리 음성 인터페이스 기술 연구

이 협 우, 육 동 석
고려대학교 컴퓨터·통신공학과

Distant-talking of Speech Interface for Humanoid Robots

Hyubwoo Lee, Dongsuk Yook
Department of Computer and Communication Engineering, Korea University
E-mail : hwlee@voice.korea.ac.kr, yook@voice.korea.ac.kr

Abstract

For efficient interaction between human and robots, speech interface is a core problem especially in noisy and reverberant conditions. This paper analyzes main issues of spoken language interface for humanoid robots, such as sound source localization, voice activity detection, and speaker recognition.

I. 서론

인간의 편의를 더욱 추구하기 위한 휴머노이드 로봇의 개발과 관심이 증가하면서 로봇의 인터페이스 개발이 주목받고 있다. 인간의 가장 편한 통신 수단인 음성을 이용하여 로봇과 상호작용을 위한 많은 연구가 진행되고 있다. 이런 연구들은 부분적으로 좋은 성과를 이루고 있지만 시스템 전반에 걸친 성공은 만족스럽지 않은 상태이며, 개선이 필요하다.

실내 환경에서 휴머노이드 로봇은 사용자가 로봇을 불렀을 때, 어느 위치에서 누가 호출하였는지를 알 수 있어야지만 비로소 인간과 원활한 의사소통을 할 수 있다. 본 논문에서는 휴머노이드 로봇 인터페이스에 필요한 음성 위치 추적 기술, 음성 구간 검출 기술, 화자 인식 기술에 대하여 알아보도록 한다.

II. 음원 위치 추적

휴머노이드 로봇이 발화자의 위치를 예측하기 위한 방법은 다채널 마이크로폰 배열을 기준으로 크게 subspace based 방법, time delay estimation (TDE) 방법, beam forming 방법 등 세 부류로 나눌 수 있다.

대표적인 subspace를 이용한 방법은 multiple signal classification (MUSIC)이다[2]. 입력된 신호에 원본 음성과 관련이 없는 잡음이 있다고 가정하고, cross correlation matrix에 eigen decomposotion을 적용하여 음성과 잡음을 나눈다. 분리한 각각은 수직의 다른 공간을 형성하고, 식 (1)처럼 표현한다.

$$P(e^{j\omega}) = \frac{1}{\sum_{i=p+1}^{N} |e^H v_i|^2} \quad (1)$$

여기서 p는 signal의 개수 이고, e^H는 signal eigen vectors의 Hermitian form이다. v_i는 noise eigen vector를 나타내며 가장 큰 $P(e^{j\omega})$ 값이 음원의 방향을 나타낸다. MUSIC은 높은 해상도를 보이지만 반향과 잡음이 있는 환경에서 다른 방법에 비해 낮은 성능을 보인다.

TDE방법 중 널리 사용되는 것으로 generalized cross correlation (GCC)이 있다[1]. 두 마이크로부터 입력받은 신호 사이의 time delay(τ)를 cross correlation을 이용하여 구하고, 이를 이용해 음원의 위치를 예측하는 방법으로 아래와 같이 나타낼 수 있다.

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_{12}(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega\tau} d\omega \quad (2)$$

(2)에서 $\Psi_{12}(\omega)$는 가중치 함수를 나타내고, $X_i(\omega)$는 i번째 입력 신호의 Fourier Transform을 나타내며 '$*$'는 complex conjugate를 나타낸다. 계산량이 적지만, 반향에 따른 현저한 성능의 저하가 단점이다.

빔형성기를 이용하는 대표적인 방법으로 steered response power(SRP)[3] 방법이 있다. 이는 미리 예측하도록 정한 좌표에서 filter-and-sum의 출력 파워

($P(q)$)를 구하여 가장 큰 값을 갖는 지점을 소리의 발생 위치라고 예측하는 것으로 다음과 같이 표현된다.

$$P(q) = 2\pi \sum_{n=1}^{N} \sum_{m=1}^{N} R_{x_n x_m}(\tau_{mn}) \quad (3)$$

여기서 q는 빔형성기에 의해 계산된 위치 좌표이고, N은 마이크 개수, $\tau_{mn} = \tau_m - \tau_n$은 두 마이크의 time delay 차이를 나타낸다. SRP는 다른 음원 위치 추적 방법들 보다 좋은 성능을 보이지만 지정한 모든 위치에 대하여 값을 구하므로 계산량이 많다.

위의 방법들은 음원에서 마이크까지 소리가 직접 들어오는 것을 가정하므로 휴머노이드 로봇과 같이 마이크 배열 위치와 개수에 제약이 있는 상황에서는 적용하기 어렵다. 또한, 가용 마이크의 수가 줄어들면 전체 시스템에서 얻을 수 있는 signal-to-noise ratio(SNR)이 낮아짐으로써 성능저하를 야기하므로 로봇환경에 적합한 새로운 위치 추적 방법의 연구가 필요하다.

III. 음성 구간 검출

녹음된 소리에서 음성 구간을 검출하는 문제는 오래된 문제임에도 불구하고, 아직 지속적인 연구가 필요한 분야이다. 널리 알려진 방법인 입력음의 에너지나 zero-crossing rate(ZCR)을 이용할 경우, 잡음의 에너지가 음성의 에너지 보다 큰 많은 경우에 있어서 성능의 악화를 초래한다. 이와 같은 낮은 SNR 환경에서의 단점을 보안하기 위해 hidden markov model(HMM)을 이용하거나, 엔트로피를 이용하는 등의 많은 음성 구간 검출연구가 이루어져 왔다[4][5].

하지만 실내 생활 잡음과 로봇 자체의 복합적인 잡음이 있는 환경의 휴머노이드 로봇 시스템에 적용하기에는 안정적인 판단기준을 제시하지 못하고, 잘못된 결과를 보인다. 그러므로 실내 환경에 강인한 음성구간 검출 연구가 필요하다.

IV. 화자 인식

휴머노이드 로봇이 미리 등록된 화자를 인식함으로써 보다 인간 친화적인 상호 작용이 가능하다. 이를 위해 등록된 화자 중 어느 누구인지를 구분하는 화자 식별(speaker identification)과 원하는 화자인지 아닌지 구분하는 화자 인증(speaker verification)등의 화자 인식 기술을 이용한다[6]. 또한, 로봇을 호출할 경우 일반적으로 원거리이므로 원거리 화자 인식 기술이 필요하며 화자가 말한 음성이 미리 약속된 특정 키워드가 아닌 자유롭게 말하는 문장 독립형(text independent) 시스템을 사용한다면 휴머노이드 로봇은 더욱 자연스러운 인터페이스를 제공할 수 있다. 동시에, 각종 생활 잡음에 대하여 반응하지 않도록 음성이외의 소리에 대한 거절 기술이 수반되어야 한다.

V. 결론

휴머노이드 로봇이 인간에게 편리성을 제공하기 위한 적절한 인터페이스의 구축은 매우 중요한 문제이다. 이런 인터페이스는 음원 위치 추적, 음성 구간 검출, 화자인식과 같은 문제에 직접적으로 연관이 되어 있으며 로봇에 적용하기 위해서는 각각의 특성에 부합하는 새로운 연구가 필요하다.

감사의 글

이 논문은 2006년도 정부재원(교육인적자원부-학술연구조성사업비)으로 학술진흥재단의 지원을 받아 연구되었음(KRF-2006-311-D0082)

참고문헌

[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.24, pp.320-327, 1976

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol.34, pp.276-280, 1986

[3] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. Thesis, BrownUniversity, 2000

[4] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian mode," *IEEE Transactions on Speech and Audio Processing*, vol.11, pp.498-505, 2003

[5] B. Wu and K. Wang, "Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments," *IEEE Transactions on Speech and Audio Processing*, vol.13, pp.762-775, 2005

[6] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of IEEE*, vol.85, pp.1437-1462, 1997